DOCUMENT RESUME

ED 376 204 TM 022 315

AUTHOR Carlson, Janet F.
TITLE Remodeling Our View of Assessment: The Test Giver as Instrument.
PUB DATE Mar 94
NOTE 9p.; Paper presented at the Annual Meeting of the National Association of School Psychologists (Seattle, WA, March 1994).
PUB TYPE Viewpoints (Opinion/Position Papers, Essays, etc.) (120) -- Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Educational Assessment; *Examiners; Experience; *Individual Differences; *Measurement Techniques; *Psychometrics; Scoring; Test Bias; Test Construction; *Testing; Test Reliability; Test Validity; Training

ABSTRACT

It is generally assumed that test administrators are accurate and dependable, and that the psychometric properties of validity and reliability applied to test givers are at acceptably high levels. The test giver is thought to have been standardized through training reinforced by experience. This paper considers validity and reliability in relation to test givers regarded as instruments of measurement. The test giver may be regarded as part of the instrument he or she uses, or the giver may be seen as the master instrument in charge of the others. It must be acknowledged that there are differences among test administrators. To ensure the best assessment by the test giver as instrument of assessment, the following must be addressed: (1) acknowledging that the giver is a person; (2) not reviewing a child's records before the assessment; (3) referring to other reports before drafting one's own; (4) talking to other test givers regularly, particularly about scoring; (5) providing training on the issue of behavioral observations; and (6) ensuring that test developers acknowledge the role of the test giver. Two tables illustrate the discussion. (Contains 38 references.) (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
JANET F. CARLSON
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED376204

Remodeling Our View of Assessment: The Test Giver as Instrument

Janet F. Carlson

The article invokes a literal image of Test Givers as measurement devices, and explores the psychometric properties of these instruments. Criterion and content validity are described, followed by test-retest, parallel forms, and internal consistency reliability. Recommendations for improving Test Givers' psychometric properties are offered.

In many ways, we assume that those who administer tests are accurate and dependable. In essence, we assume the psychometric properties of validity and reliability, as applied to Test Givers, are at acceptably high levels. For the most part, these properties are thought to have been established largely by virtue of one's graduate training. In addition, incremental gains in the Test Giver's accuracy and consistency are assumed to occur during one's internship, professional experiences, continuing educational experiences, and on-going exposure to supervision such as might occur in peer review processes.

Virtually all graduate programs that train Test Givers attempt to make them uniform. Those who have taught in graduate-degree programs and, perhaps, have taught courses in testing will recognize that trainers do not encourage diversity when it comes to learning to administer a standardized test. In fact, we emphasize the opposite--uniformity. There is a sense that Test Givers should be interchangeable. In essence, the Test Giver is thought to have been "standardized" during his or her training, and this standardization is reinforced in subsequent experiences.

However, given the variability of training programs, and the variability of instructors in courses relating to test giving, it seems unreasonable to expect such a high degree of sameness among Test Givers. Uniformity may characterize Test Givers in a particular program or in a particular course, but it is unlikely to hold across instructors and across graduate programs. In this light, let us reflect upon the psychometric properties of validity and reliability, by framing these formal issues in relation to the instruments under consideration now--Test Givers.

Validity

In short, the validity of an instrument reflects the extent to which it measures what it is intended to measure. Most of our graduate programs trained us well in terms of gathering information that answers the referral question. But just how does a Test Giver become accurate or valid in the first place? Most likely, the process is initiated during specific graduate courses and training experiences. So the beginning Test Giver takes a course in administering intelligence tests, and is instructed on how to do this accurately. Achieving accuracy is often equated with rigorous instruction on the "how to's" of test administration, scoring, and interpretation.

But many Test Givers who conduct assessments as part of their everyday professional activities have encountered unique answers to test questions--answers that do not appear (even remotely) in the test manual. Even after one's administration and scoring has been "standardized" by graduate training, such occurrences are not uncommon. In an attempt to limit the impact of such events, graduate training tends to emphasize the overarching principles of scoring over the specific responses. So students learn to score by considering the general guidelines rather than the actual words or phrases used by a test taker to answer a given question. However, even straightforward subtests, such as Information, can be problematic when they produce curious responses. Theoretically in this subtest, as in several others, there is one correct response to each question, with an occasional second or third option that receives credit as well. Follow-up prompts or

---

This paper was presented in March of 1994 at the Annual Convention of the National Association of School Psychologists in Seattle, Washington.

Janet F. Carlson is an assistant professor of Counseling and Psychological Services at the State University of New York, College at Oswego. Correspondence concerning this paper should be addressed to Janet F. Carlson, CPS Department, Mahar Hall, SUNY-Oswego, Oswego, NY 13126.

TM022315

BEST COPY AVAILABLE

questions are delineated clearly in the test manual. But many of us would question what to say in response to the child who says "Fred" when asked, "What do we call a baby cow?" We might be especially interested in probing the response, or perhaps even giving credit for it, if the response came from an inner city child of limited means, who has had little experience with rural terms for baby farm animals.

Many graduate programs emphasize the administration, scoring, and interpretation of tests. Most also attempt to address the underlying constructs at issue--constructs such as intelligence. Of the many things traditionally covered in courses on test giving, administration and scoring appear to be the most simple, task-oriented components of the process. So a logical question might be: How successful are graduate programs in teaching these sorts of skills?

Many studies have demonstrated that graduate students and professionals alike commit numerous errors both in administration and in scoring of standardized assessment instruments (e.g., Blakely, Fantuzzo, Gorsuch, & Moon, 1987; Brannigan, 1975; Conner & Woodall, 1983; Franklin, Stillman, Burpeau, & Sabers, 1982; Hanna, Bradley, & Holen, 1981; Miller & Chansky, 1972; Moon, Blakely, Gorsuch, & Fantuzzo, 1991; Moon, Fantuzzo, & Gorsuch, 1986; Sherrets, Gard, & Langner, 1979; Slate & Chick, 1989; Thompson & Bulow, 1994; Warren & Brown, 1972). Much of the research in this area has focused upon intelligence tests, and the implications have addressed such factors as accuracy and the effects on placement decisions that follow from such mistakes. Slate and Hunnicutt (1988) reviewed the literature on Wechsler scoring errors and suggested several factors that might account for the departures that are so widely noted, including carelessness on the part of the Test Giver and poor instruction on the part of the trainer.

Table 1
WISC-R Subtest and Composite Scores Assigned by Different Scorers for Same Subject

	Verbal							Performance				Composites		
Scorer	Inf	Sim	Ari	Voc	Com	DSp	PC	PA	BD	OA	Cdg	Mz	VIQ	PIQ	FSIQ
1	12	15	10	13	13	10	14	14	13	13	8	12	115	117	118
2	12	19	10	13	14	7	14	14	13	11	8	7	122	114	121
3	12	15	10	13	13	10	14	14	13	11	8	12	115	114	117
4	12	15	10	12	12	10	9	14	13	11	8	8	113	106	111
5	12	15	10	14	14	10	13	14	13	11	8	10	118	112	118
6	12	18	10	12	13	10	14	14	13	11	8	12	118	114	118
7	12	16	10	13	13	10	14	14	13	11	8	12	117	114	118
8	12	18	10	13	13	10	14	14	14	11	8	12	118	115	120
9	12	19	10	12	14	9	14	14	13	11	8	12	120	114	120
10	12	16	10	13	13	10	14	14	13	11	8	12	117	114	118
11	12	16	10	12	12	5	14	14	13	11	8	12	114	114	116
12	12	16	10	12	14	10	14	14	13	11	8	12	117	114	118
13	12	16	10	12	13	10	14	14	13	11	8	12	115	114	117
14	12	16	10	12	13	4	14	14	13	11	8	12	115	114	117
15	12	17	10	12	14	10	14	14	14	11	8	12	118	115	119
Mean	12.0	16.5	10.0	12.5	13.2	9.0	13.6	14.0	13.1	11.1	8.0	11.3	116.8	113.7	117.7
s.d.	0.0	1.41	0.0	0.64	0.68	2.0	1.30	0.0	0.35	0.52	0.0	1.62	2.37	2.35	2.28
Expert	12	15	10	13	13	10	14	14	13	11	8	12	115	114	117
% agr	100	26.7	100	40.0	53.3	73.3	86.7	100	86.7	93.3	100	80.0	26.7	66.7	20.0
%±1	100	66.7	100	100	100	80.0	93.3	100	100	93.3	100	80.0	33.3	80.0	66.7

In a subsequent empirical article, Slate, Jones, and Murray (1991) note that practice administrations of Wechsler tests merely permitted students in training to practice their errors rather than to improve their proficiency. They also note, as have others, that Verbal subtests are particularly prone to examiner errors. Predictably, the most frequent number of errors on Wechsler tests occurred for Vocabulary, Comprehension, and Similarities, followed by Picture Completion and Information. The ten most common types of errors made were (1) a failure to record something (response or time), (2) assigning too many points, (3) a failure to question a response, (4) questioning inappropriately, (5) assigning too few points, (6) incorrect conversion of raw scores to standard score, (7) failure to obtain a "ceiling", (8) failure to assign points correctly on Performance items, (9) incorrect raw score for subtest total--a math error, and (10) incorrect calculation of chronological age--another math error.

Table 2
Sample of Behavioral Observations Made by Different Scorers for Same Subject

Scorer	Observations
3	V. remained attentive throughout the testing procedures. She was cooperative in answering test items but offered little spontaneous speech. V. appeared concerned with her performance. She was especially persistent and worked deliberately on the performance subtests. On some of the verbal subtest items, she clearly stated her lack of knowledge (e.g., "We don't study things like that," "I have *no* idea"). Throughout testing, V. was frequently moving her left foot underneath the table.
8	V. did not fidget in her chair. She seemed confortable and often smiled at the examiner revealing that there was a nice rapport established. She also helped the examiner in some cases which also showed her comfortness and patience. V. had little to say throughout the test and appeared to be quite confident in her answers, and sure that she had never heard some of the words before and therefore just did not know the answers to them.
10	V. was very cooperaative during testing. She helped put the materials away on object assembly. She followed directions. She was also rather quiet when not asked for a response. V. was clicking her heels during a few of the subtests. In digit span, she mouthed the numbers and was playing with her hair. V. tended to fix the cards in picture arrangement not the blocks in block design when she was finished.
11	V. seemed anxious at times. Her foot was rocking underneath the table a lot during testing. Also she would cover her face, giggle and lookd down when she didn't know something, especially the verbal subtests. However, generally, she seemd confident and enjoyed the tasks. She concentrated well and was careful with her work. She was also pleasant and cooperative. She stated that the mazes and blocks and a few of the puzzles were hard for her.
13	V. is a white girl of average height and weight. She appeared neat and well groomed at the time of testing. V. spoke sofly throughout the testing and spoke with moderate affect. During testing, she sat with her arms folded in front of her on the talbe. She also sat up straight in her chair without slouching. While V. showed little undue anxiety, she did demonstrate that she was taking the testing situation seriously, as described by her behaviors above. She worked quietly on most tasks with little verbalization other than what she was asked to contribute. Overall, she was cooperative and followed directions given to her.

In a small-scale empirical investigation of my own, I examined beginning examiners' scoring errors. My exploration differed from others in that I limited the scope to include *only* scoring errors by presenting students with a videotaped administration of a WISC-R. Thus, all the students had to do was record and score responses. Because they did not need to be concerned with querying, setting up test materials, or obtaining proper basals and ceilings, they could not make these kinds of mistakes. None of the 15 students in this small study made math errors, but they did make all the rest. The numerical results of this study are displayed in Table 1.

Also of considerable concern is what has been left out in research of this kind. Traditionally, the research has looked at score accuracy somewhat and at competence in administration. Behavioral observations are notoriously absent from consideration, with but a few exceptions (e.g., Glutting, Oakland, & McDermott, 1989; Kaplan, 1991). It seems another assumption is made regarding Test Givers--they have an innate capacity to observe behavior accurately and need little instruction or guidance on these tasks.

Certainly, it cannot be that behavioral observations are unimportant. Indeed, authors on this subject nearly always note that "behavioral observations" are part of the formal report (Ownby, 1987; Ross-Reynolds, 1990; Tallent, 1983; Zuckerman, 1989). Some have made specific suggestions about which behaviors to include in this section of the report. One practitioner and internship supervisor I know routinely notes that the behavioral observations section of a formal report is both the most important section and the most difficult section to write.

Just how accurate are Test Givers in their observations of behavior? Table 2 contains some of the behavioral observations made by students about the videotaped subject used in my study. Although the descriptions are all of the same child, the differences are apparent, and leave us questioning how to improve the accuracy and/or standardization of this important part of assessment. Errors made in observing behavior may be more difficult to address than errors of administration and scoring, because they are more vague, more elusive and more open to subjective interference. And although some research has been directed at identifying scoring and administration errors, relatively little has appeared with a focus on behavioral observations, as noted earlier.

A view of **criterion validity** can be had by considering the information presented in Table 1. If each individual Test Giver's scoring pattern is compared to that produced by the panel of the experts, the comparison yields a kind of criterion validity. Here, the accepted criterion measure to which individual results (i.e., the scores assigned by individual Test Givers) are compared is the score profile produced by the expert panel. The extent to which individual Test Giver's scores agree with those of the expert panel yields a measure of criterion validity for each individual Test Giver.

Alternatively, Test Givers can be imagined as a group of items, with each item representing an individual Test Giver. That is, it is possible to think of all Test Givers collectively as a single instrument, because we shape Test Givers in groups and try to train good "troops" of psychologists or counselors or whatever, as far as their test giving is concerned. At least for the ensuing discussion regarding **content validation**, it is helpful to think in these terms.

One obvious question that follows from this view of Test Givers and the content validity issue under discussion has to do with the adequacy of the sample. If Test Givers are seen collectively as an instrument of assessment, and if each individual Test Giver is regarded as an item of that instrument, one might question how well the items represent the domain of interest. Arguably, the domain of interest is "human beings" or, more narrowly for our purposes, citizens of the United States. Clearly, the domain should be inclusive rather than exclusive, as it would not be desirable to have the collection of Test Givers exclude, in whole or in part, identifiable segments of the general populous. So the content validity question becomes: To what extent does the collection of Test Givers reflect the U.S. citizenry? The answer: To a limited extent, at best, given what is known about the demographic characteristics of the various professions concerned with assessment. The underrepresentation of most minority groups, bilingual persons, and persons from lower socioeconomic and disadvantaged backgrounds speak to this issue. We must admit that content validity, as defined above, is weak.

Relatedly, questions may be raised about the process of item development. That is, thinking as one does when developing a traditional instrument, a test developer must be concerned with the characteristics of the pool of items. The process is analogous to that of traditional item development. We start with a pool of items--more than we plan to retain. Items are eliminated on the basis of poor performance. In the same way, graduate schools start with more potential Test Givers than will actually complete their training. Along the way, some of these are eliminated for reasons including poor performance. Most graduate schools have attempted to improve the

representativeness of the item pool by active efforts to recruit and retain underrepresented Test Givers. Some have succeeded in these efforts; others have attempted to remedy pool problems by providing additional training or requiring additional course work in multiculturalism or multicultural service delivery and so forth.

Reliability

Reliability refers to the dependability of test scores; that is, their consistency. Essentially, reliability reflects the confidence one can have that test scores will remain the same across time, across persons (that is, scorers), across versions of the same instrument, and across portions of the test itself. If Test Givers are instruments, too, then they are expected to be reliable. That is, it seems reasonable to examine the consistency of their scores.

It is possible to design studies to explore the consistency of scores assigned by Test Givers across time. Doing so would correspond to a determination of **test-retest reliability**. To do this, one might have Test Givers assign scores to a number of test responses, wait a while, and have them do it again. To my knowledge, such a study has not been conducted.

A similar approach could be used to explore Test Givers' observations of a test taker's behavior during the assessment process. This kind of research might involve videotaping test administrations, showing the tapes to several Test Givers and having them rate the test taker's behavior on a behavioral rating scale at two different points in time. They could also draft a few paragraphs describing each test taker's behavior, and a number of judges could then evaluate the similarities of the descriptive paragraphs.

If we once again invoke the image of the Test Givers as independent instruments, a kind of reliability estimate that mimics **parallel** or **alternate forms** takes shape. Each Test Giver is, after all, thought to be interchangeable and so imagining them all as parallel is not difficult. Some research has appeared along these lines (e.g., Kasper, Throne, & Schulman, 1968; Oakland, Lee, & Axelrad, 1975). The data in Table 1 also can be examined in light of this kind of reliability. The instruments (i.e., the Test Givers) all saw the same video and heard the same responses and had very similar instruction. Theoretically, they should have arrived at the same scores.

One could also view the Test Givers collectively, as a single instrument as previously suggested. If this were the case, then a kind of **internal consistency** measure could be approximated. For example, a rudimentary split-half reliability might be accomplished by using an odd/even split and computing the correlation coefficient between the two halves. Even without performing the calculation, one can see that the internal consistency of this instrument (the collective group of Test Givers) is quite high. Some research investigating consistency of scoring can be viewed as addressing internal consistency (e.g., Bradley, Hanna, & Lucas, 1980; Miller, Chansky, & Gredler, 1970; Ryan, Prifitera, & Powers, 1983).

Factors Influencing Test Performance

In numerous studies, a variety of factors have been suggested to influence the test *taker's* performance. For example, performance on intelligence tests has been explored in relation to such factors as the Test *Giver's* sex, age, ethnicity, socioeconomic status, training and experience, appearance, and personality characteristics (Anastasi, 1988, p. 38). Significant findings have emerged for all of these factors at one time or another, but the soundness of some of these studies puts their findings in question. Still, these types of investigations raise questions about how some of these same factors may affect the Test Giver or procedures used by him or her during test administration and scoring (Geisinger & Carlson, in press). That is, when these factors are going in the other direction--when they emanate from the test taker to the Test Giver, what effects, if any, occur?

We should consider that assessments of students from bilingual or culturally diverse backgrounds may need to break with tradition somewhat (Rogers, 1993; Rosado, 1986). Similarly, decisions emanating from these assessment may need to proceed in a manner that takes into account environmental factors (Reynolds & Kaiser, 1990). Perhaps these assessments and decisions will need to make greater use of observation and judgment and less use of standardized instruments. In keeping with this idea, Figueroa (1990) states that in conducting assessments of such students, there is "no reason to assume that...a judgment call will contain more error than a psychometric test." But no one said it will contain less either. At the very least, Test Givers need to bear in mind what effects culture may have not only on the test taker's behavior and performance, but also on the Test Giver's interpretation of scores and especially of test behavior (American Psychological Association, 1991; Dana, 1993; Miller-Jones, 1989; Ogbu, 1988). For example, the Test Giver should *not* draw the same conclusions for all test takers who do not make eye contact or do not converse readily with him or her, as these factors are likely to be shaped differently by different cultures. The Test Giver must be aware of the many influences that a

particular culture or subculture may have on a child's behavior (Geisinger, 1992).

Conclusions and Recommendations

Perhaps it is best to think of Test Givers as part of each instrument they use, rather than as separate instruments entirely. Or, we could view the Test Giver as the "master" instrument in charge of the others.

If we adopt the first position--that the Test Giver becomes a part of whatever instrument he or she uses--then we would have to acknowledge that when we administer what is considered to be a standardized instrument, there is one component of that instrument that does not stay the same. It would be comparable to opening a WISC or Stanford-Binet kit and having a new subtest each time, or at least several new items.

Indeed, during the standardization of most tests, test developers use or train experts to administer the test, and review these administrations to eliminate differences in procedures or interpretations. Typically, these differences are eliminated in advance of the norming process. In doing so, the test developers are trying to override the person-to-person differences inherent in Test Givers who--after all is said and done--are people. In a very real sense, this procedure bespeaks the role of the Test Giver as a part of the instrument. With the kind of close scrutiny that is given to these Test Givers, differences stemming from person-to-person variations are considerably reduced.

Of course, most Test Givers do not have the benefit of such close scrutiny and validity checks once they have completed graduate school. When a new edition of a test is developed, some professionals find it necessary or desirable to attend training workshops in order to be updated on the changes and procedural implications of the changes. But many times, this does not seem necessary or is not within the budget, or is impossible for some other reason, and psychologists end up teaching themselves the revised edition (Chattin & Bracken, 1989; Dumont & Faro, 1993).

Given that Test Givers are part of the assessment and might even be considered as instruments themselves, what can we do--in light of the foregoing--to be better at the task of assessment? What can be done to improve the psychometric properties of these instruments? A few suggestions are offered below.

1. **Acknowledge that you are a person**. You have traits--physical and personal ones--that enter into the assessment process, no matter how rigorously your training program tried to obliterate these. So during assessments, take note of how test takers interact with you. That is, note how test takers typically respond to you--do they view you as the "enemy", as a "friend", as a "parent", as someone who is going to "uncover a secret"? If so, then *you* are bringing that to the testing situation. And when the test taker's behavior reflects this trait of yours, you must see the behavior *not* as something that belongs to the test takers so much as belonging to you.

2. **Do not review test results of a child referred for reevaluation before conducting the assessment**. Although doing so creates an appearance of reliability, this form of reliability concerns the inanimate instruments primarily and not the Test Givers. At best, it supports the reliability of the assessment procedure *sans* the Test Giver.

3. After collecting the assessment information and drafting the report, **refer to the other report/s before finalizing yours**. Doing so can--but will not necessarily--serve as a validity check. Of course, one would not expect identical assessments to emerge, but making use of information in the record will highlight changes that have occurred and might also indicate those areas in which some double-checking might be in order. Address differences that appear, hopefully in light of changes in the subject, rather than errors in the instruments (that is, the Test Givers).

4. **Talk to other instruments** (i.e., Test Givers) regularly. Discuss the scoring of specific items. Consider exchanging record forms with your colleagues from time to time in order to check consistency of your scores. Look for a pattern...do you score low? high? Are there particular types of tests, subtests or items where you tend to differ? Resolve those differences to the maximum extent possible.

5. For those who train Test Givers, **consider including some training on the issue of behavioral observations**, and some activity that relies upon consensus, such as viewing the same test administration, scoring it, and writing behavioral observations.

6. **For test developers, acknowledge the role of the Test Giver in a forthright manner**. Include in the test manual reports on typical variations in scores, and identify the factors that contribute to these variations. They are not all random errors. At the time of test standardization, test developers might also sponsor or design research to address the role of the Test Giver.

References

American Psychological Association (1991). *Guidelines for providers of psychological services to*

ethnic and culturally diverse populations. Washington, DC: APA (Division 16).

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Blakely, W., Fantuzzo, J., Gorsuch, R., & Moon, G. (1987). A peer-mediated, competency-based training package for administering and scoring the WAIS-R. *Professional Psychology: Research and Practice, 18,* 17-20.

Bradley, F. O., Hanna, G. S., & Lucas, B. A. (1980). The reliability of scoring the WISC-R. *Journal of Consulting and Clinical Psychology, 48,* 530-531.

Brannigan, G. G. (1975). Scoring difficulties on the Wechsler intelligence scales. *Psychology in the Schools, 12,* 313-314.

Chattin, S. H., & Bracken, B. (1989). School psychologists' evaluation of the K-ABC, McCarthy Scales, Stanford-Binet IV, and the WISC-R. *Journal of Psychoeducational Assessment, 7,* 112-130.

Conner, R., & Woodall, F. E. (1983). The effects of experience and structured feedback on WISC-R error rates made by student-examiners. *Psychology in the Schools, 20,* 376-379.

Dana, R. H. (1993). *Multicultural assessment perspectives for professional psychology.* Boston: Allyn & Bacon.

Dumont, R., & Faro, C. (May 1993). The WISC-III: Almost two years old; proceeding with caution--Practitioners' concerns. *NASP Communiqué,* 12-15.

Figueroa, R. A. (1990). Best practices in the assessment of bilingual children. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology-II* (pp. 93-106). Washington, DC: National Association of School Psychologists.

Franklin, M. R., Jr., Stillman, P. L., Burpeau, M. Y., & Sabers, D. L. (1982). Examiner error in intelligence testing: Are you a source? *Psychology in the Schools, 19,* 563-569.

Geisinger, K. F. (Ed.) (1992). *Psychological testing of Hispanics.* Washington, DC: American Psychological Association.

Geisinger, K. F., & Carlson, J. F. (in press). Standards and standardization. In J. Butcher (Ed.), *Practical considerations in clinical personality assessment.* New York: Oxford University Press.

Glutting, J. J., Oakland, T., & McDermott, P. A. (1989). Observing child behavior during testing: Constructs, validity, and situational generality. *Journal of School Psychology, 27,* 155-164.

Hanna, G. S., Bradley, F. O., & Holen, M. C. (1981). Estimating major sources of measurement error in individual intelligence scales: Taking our heads out of the sand. *Journal of School Psychology, 19,* 370-376,

Kaplan, C. (1991, August). Observing behavior during testing: Development of a method for quantifying clinical judgment. Paper presented at the ninety-ninth annual convention of the American Psychological Association, San Francisco, CA.

Kasper, J., Throne, F., & Schulman, J. (1968). A study of the inter-judge reliability of the responses of a group of mentally retarded boys to three WISC subscales. *Educational and Psychological Measurement, 28,* 469-477.

Miller, C. K., & Chansky, N. M. (1972). Psychologists' scoring of WISC-R protocols. *Psychology in the Schools, 9,* 144-152.

Miller, C. K., Chansky, N. M., & Gredler, G. R. (1970). Rater agreement on WISC protocols. *Psychology in the Schools, 7,* 190-193.

Miller-Jones, D. (1989). Culture and testing. *American Psychologist, 44,* 360-366.

Moon, G., Blakely, W., Gorsuch, R., & Fantuzzo, J. (1991). Frequent WAIS-R administration errors: An ignored source of inaccurate measurement. *Professional Psychology: Research and Practice, 22,* 256-258.

Moon, G., Fantuzzo, J., & Gorsuch, R. (1986). Teaching WAIS-R administration skills: Comparison of the MASTERY model to other existing clinical training modalities. *Professional Psychology: Research and Practice, 17,* 31-35.

Oakland, T., Lee, S. W., & Axelrad, K. M. (1975). Examiner differences on actual WISC protocols. *Journal of School Psychology, 13,* 227-233.

Ogbu, J. U. (1988). Human intelligence testing: A cultural-ecological perspective. *National Forum: Beyond Intelligence Testing, 68* (2), 23-29.

Ownby, R. L. (1987). *Psychological Reports.* Brandon, VT: Clinical Psychology Publishing Co.

Reynolds, C. R., & Kaiser, S. M. (1990). Test bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology* (2nd ed.) (pp. 487-525). New York: Wiley.

Rogers, M. R. (1993). Psychoeducational assessment of racial/ethnic minority children and youth. In H. B. Vance (Ed.), *Best practices in assessment for school and clinical settings* (pp. 399-440). Brandon, VT: Clinical Psychology Publishing.

Rosado, J. W. (1986). Toward an interfacing of Hispanic cultural variables with school psychology service delivery systems. *Professional Psychology: Research and Practice, 17,* 191-199.

Ross-Reynolds, G. (1990). Best practices in report writing. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology-II* (pp. 621-633). Washington, DC: National Association of School Psychologists.

Ryan, J. J., Prifitera, A., & Powers, L. (1983). Scoring reliability on the WAIS-R. *Journal of Consulting and Clinical Psychology, 51,* 149-150.

Sherrets, S., Gard, G., & Langner, H. (1979). Frequency of clerical errors on WISC protocols. *Psychology in the Schools, 16,* 495-496.

Slate, J. R., & Chick, D. (1989). WISC-R examiner errors: Cause for concern. *Psychology in the Schools, 26,* 78-83.

Slate, J. R., & Hunnicutt, L. C., Jr. (1988). Examiner errors on the Wechsler scales. *Journal of Psychoeducational Assessment, 6,* 280-288.

Slate, J. R., Jones, C. H., & Murray, R. A. (1991). Teaching administration and scoring of the Wechsler Adult Intelligence Scale-Revised: An empirical evaluation of practice administrations. *Professional Psychology: Research and Practice, 22,* 375-379.

Tallent, N. (1983). *Psychological report writing* (2nd ed.). Englewood Cliffs, NJ: Prentice-Hall.

Thompson, A. P., & Bulow, C. A. (1994). Administration error in presenting the WAIS-R blocks: Approximating the impact of scrambled presentations. *Professional Psychology: Research and Practice, 25,* 89-91.

Warren, S. A., & Brown, W. G. (1972). Examiner scoring errors on individual tests. *Psychology in the Schools, 10,* 118-122.

Zuckerman, E. L. (1989). *The clinician's thesaurus: A guidebook for wording psychology reports and other evaluations.* Pittsburgh, PA: Three Wishes Press.